CDF/MEMO/STATISTICS/PUBLIC/6576
Version 1

Benefits of Blind Analysis Techniques

Joel G. Heinrich—University of Pennsylvania

July 17, 2003

Abstract

Blind analysis strategies, designed to minimize the possibility of bias in experimental results, have become increasingly popular for high energy analyses in recent years. In this note, experimenter bias is discussed, motivations for blind analysis are presented, and basic techniques are described.

1 Introduction

The use of blind techniques in mainstream science began in the 1930's, when the "double blind" procedure started to gain popularity in medical research. The Oxford English Dictionary[1] defines "double blind" as:

> Applied to a test or experiment conducted by one person on another in which information about the test that may lead to bias in the results is concealed from both the tester and the subject until after the test is made; originally used of tests for determining the efficacy of drugs.

and gives examples of usage starting in 1937:

> The data consisted of the patients' judgments regarding changes in pain. These data were secured in a manner relatively free of bias by the use of the 'blind test'. *1937 JAMA 26 June 2178/2*

> The study was conducted by the 'blind' method. The materials for injection were unknown to the observer as well as to the subject. *1948 Amer. Heart Jrnl. XXXVI 529*

> The 'internal-evaluation' was made by skilled questioning under conditions of the 'double blind test' in which neither the physician nor the patient knew at the time whether the evaluation related to the placebo or khellin. *1950 Amer. Jrnl. Med. IX 146/1*

> The only safe way to obtain unbiased opinions from either of them [doctor or patient] is to make them express their opinions without knowing whether the patient received an active drug or not. This is known in America as a double blind test. *1954 Proc. R. Soc. Med. XLVII 197*

> Statistics and certain concepts, such as double-blind trials, are on everyone's mind to-day. *1961 Lancet 19 Aug. 423/1*

Double blind procedures are now universally taught and employed in the fields that experiment on human subjects: clinical research and psychology. It has long been recognized in those fields that, to avoid experimenter bias, it is not sufficient just to hide certain information from the human subjects; the information must also be hidden from the experimenters by the experimenters themselves until after the analysis is complete.

In the physical sciences, the benefits of using blind techniques to reduce bias have been less often discussed until quite recently. Nevertheless, there is early evidence of concern about experimenter bias in particle physics in this statement by Rutherford:

> It seems to me that in some way it is regrettable that we had a theory of the positive electron before the beginning of the experiments. Blackett did everything possible not to be influenced by the theory, but the way of anticipating results must inevitably be influenced to some extent by the theory. I would have liked it better if the theory had arrived after the experimental facts had been established. *Ernest Rutherford*[2]

Here Rutherford refers to Patrick Blackett, who won a 1948 Nobel prize for his 1930's experiment that used a cloud chamber to identify positrons in cosmic ray air showers. Theorists, beginning with Dirac, had already predicted the e^+, and Rutherford (1933) is obviously aware that this knowledge of what to expect can easily bias the experiment. Rutherford's phrase "everything possible not to be influenced by the theory" translates to our "blind procedure", and his "influence" translates to our "experimenter bias".

A more recent example is associated with the search for fractional charge in ordinary matter by a group of Stanford physicists led by William Fairbank. Initially, in a series of results published 1977–81, the group claimed "unambiguously the existence of fractional charges of $\frac{1}{3}$e"[3]. Luis Alvarez subsequently proposed that "blind tests" be employed[4], in which a randomly

chosen charge, of value unknown to the experimenters, would be added to the data, thereby hiding the answer until the analysis was complete. Subsequent measurements[5] by the Stanford group did incorporate the blind test: these did not confirm the original "discovery".

2 Statistical Bias

There is a huge literature concerning experimenter bias—a google search for the exact phrase "experimenter bias" locates ∼2200 web documents, and (the synonymous) "experimenter effect(s)" yields another 2000. But only a very small fraction of it is written by or for physicists. Before we discuss experimenter bias, we need a definition of (general statistical) bias[1], which we quote from reference [6]:

> Bias: Let $d(X)$ be an estimator of the unknown parameter θ. The bias is defined by
>
> $$b(\theta) = \mathcal{E}_X[d(X) - \theta]$$
>
> where the expectation $\mathcal{E}_X[d]$ is with respect to an ensemble of random variables $\{X\}$. The bias is just the difference between the expectation value, *with respect to a specified ensemble*, of the estimator $d(X)$ and the value of the parameter being estimated. If the ensemble is not given, the bias is undefined. If an estimator is such that $b(\theta) = 0 \ \ \forall\theta$ then the estimator is said to be unbiased; otherwise, it is biased. Bias, in general, is a function of the unknown parameter θ and can, therefore, only be estimated. Further, bias is a property of a particular choice of metric. In high energy physics, much effort is expended to reduce bias. However, it should be noted that this is usually at the cost of increasing the variance and being further away, in the root-mean-square sense, from the true value of the parameter. See also Ensemble, Quadratic Loss Function, Re-parameterization Invariance.

The random variable X referred to in the definition is a vector that represents the data, and the "ensemble" contains all possible experimental results (i.e. all possible X). The fact that the data X come from a random distribution that depends on the unknown parameter θ means that the estimator $d(X)$, a

[1]The term "bias" is heavily overloaded in Statistics. Here we refer to "point estimate bias"—not to be confused with "test bias" or "interval bias".

function of the observed data, also has a random distribution that depends on θ.

Reference [7] offers the following instructive comment on the definition of bias:

> There is some arbitrariness in this definition, since there exist other measures of the center of a distribution which could have been used. The expectation (or arithmetic mean) is conventionally chosen for convenience, and because of its properties for the normal distribution. Note that if an estimator is unbiased, this does not necessarily mean that on the average half of the estimates will lie above θ and half below. This would have been true if the median had been chosen instead of the arithmetic mean.

This kind of bias is normally just an inconvenience: it can be corrected for if necessary. The experiment, once performed, provides us with an estimate for θ, which can be used to estimate the bias and correct the answer. For example, defining $f(\theta) = b(\theta) + \theta$, we have $f(\theta) \simeq d(X)$, so $\theta \simeq f^{-1}(d(X))$, and $f^{-1}(d(X))$ is our (approximately) bias-corrected estimate of θ. That is, $d_0(X) = f^{-1}(d(X))$ represents a new estimator (of the unknown parameter θ) that should remove most of the bias of the original estimator $d(X)$.

Mere statistical bias (as defined above) is not the problem that we are concerned with in this note. Experimenter bias occurs when human behavior enters the equation.

3 Experimenter Bias

Let's focus on the function $d(X)$ which maps the raw data X into our estimate of the parameter θ. The idea is simple enough: the detector produces an actual data sample X_0, we calculate $d(X_0)$ (a single real number in this example), and publish it, along with an explanation of $d(X)$, uncertainty estimates, etc.

But the function $d(X)$ is typically quite complicated. It may include detector calibration, reconstruction algorithms, efficiency or acceptance determination, data selection cuts, artificial neural networks, complicated empirical parametrizations of backgrounds and signal, maximum likelihood fits with dozens of free parameters, etc. It's complicated enough that Monte Carlo simulation is necessary.

The defining of $d(X)$ for use in measuring a particular physical quantity θ is the task of the experimenters. Once the complete specification of how to calculate $d(X)$ for all possible X is fixed, the methods of Statistics—frequentist or Bayesian—will be able to characterize the relationship between $d(X_0)$ and θ in terms of probabilities.

Experimenter bias is possible when the algorithm $d(X)$ is modified by the experimenter based on (partial or complete) information about the value of $d(X_0)$. In this case, the algorithm $d(X)$ really includes a biological neural network—the brain of the experimenter—since as X varies, the experimenter's decision of the form that $d(X)$ should take also can vary. Proper Monte Carlo simulation of $d(X)$ in such a case is impossible; we must include a simulation of the experimenters' decision process. That is, we need to know what the experimenters would have decided for $d(X)$ (based on their looking at the data) for all possible data X.

Bias (in the statistical sense) will then occur through unconscious feedback between the choice of the algorithm $d(X)$ and the experimenters expectation of or preference for a particular result. Since this process is unconscious, the danger is always present when the experimenters are even vaguely aware of the effect of their modifications to $d(X)$ on the difference between the result and the expected or preferred value. Because this process is unconscious, it cannot be modeled properly, and therefore the actual size of the bias can not be objectively estimated, even by the experimenters themselves. The experimenters will not even be aware that it has occurred, although they might admit that it was possible.

The following are examples of experimenter bias that might occur in a high energy physics analysis:

- The data selection criteria ("cuts") are unconsciously adjusted to bring the answer closer to a theoretical value or a previously measured value.

- Comprehensive checks are performed if the answer disagrees with expectation, otherwise not so comprehensive. The extra checks might be invented by the people actually performing the analysis, or requested by godparents, physics groups, etc. (The experimenters feel more confident when the answer comes out "right".) (These checks may lead to "corrections" that change the answer.)

- Extra systematic uncertainties are invented and attached to the answer when it disagrees with expectations.

- Several competing analyses are performed using the same data. The physics group charged with making the decision chooses which is worthy of publication *after* learning the answers, unconsciously favoring analyses that "come out right".

In each case, the experimenter bias is unintentional—the experimenters normally know that these practices are objectionable, but in each example, the course of the analysis is unconsciously influenced by their knowledge of how the outcome is affected.

Note that early knowledge of the value of an ancillary statistic should not lead to experimenter bias. Reference [6] defines an ancillary statistic as:

> Consider a probability density function $p(X|\theta)$. If the distribution of the statistic $t(X)$ is independent of θ and the statistic is also independent of θ, then the function $t(X)$ is said to be an ancillary statistic for θ. The name comes from the fact that such a statistic carries no information about θ itself, but may carry subsidiary (ancillary) information, such as information about the uncertainty of the estimate. *Example* In a series of n observations x_i, n is an ancillary statistic for θ. The independence of the distribution on the ancillary statistic suggests the possibility of inference conditional on the value of an ancillary statistic. See also Conditioning, Distribution Free, Pivotal Quantity, Sufficient Statistic.

(The example given above is, perhaps, too general, and might be confusing—"n is an ancillary statistic for θ equal to the expectation value of the x_i" is a simple example that one can easily show is ancillary.) In some cases, the size of the uncertainty (of the result $d(X_0)$) is itself ancillary.

Blind techniques are designed to minimize the experimenters' exposure to any non-ancillary statistic before the design of $d(X)$ is finalized.

4 Historical Examples of Experimenter Bias

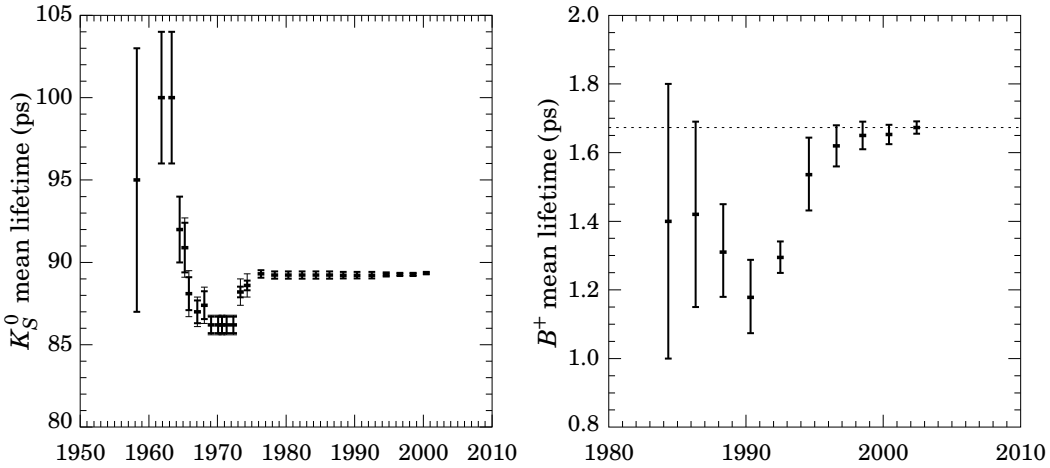
Examples of experimenter bias are not difficult to locate. One method is to look for past results that agree with expectations too well (i.e. P-value very close to 1.0). The classic case is Gregor Mendel's 1865 work on inheritance. Quoting from reference [8]:

> In every case the data agreed with the theoretical ratios within less than the standard errors; taking the whole together, χ^2 was 41.6 on

84 degrees of freedom, and the chance of a smaller value arising accidentally is 7×10^{-5}.

Since there was no inheritance theory when Mendel's work was published, Mendel (or his assistants) might have been biased by his expectations arising from his earlier studies, or perhaps just by the simple small-integer fractions involved. For example, in one class of (pea) population used in Mendel's investigations, the recessive trait of interest does occur with a probability of 1/4—once this trend was recognized in the data, there may well have been an unconscious bias toward 1/4. If the probability had been something less recognizable, say 9/31, the tendency toward bias might have been weaker. (A modern scientist would have noticed immediately that his agreement with his expectation was too good, but even this most basic statistical principle was unknown at that time.)

Another strategy is to look for past results that are too inconsistent with current established values. The particle data group provides a dozen historical plots with every "Review of Particle Physics"; we show the two lifetime plots from the most recent issue [9]:



World average K^0_S and $B^\pm$ mean lifetime vs PDG publication year.

It is important to understand that the data points show the PDG's weighted average of all measurements available[2], so the points do not change from issue to issue at all unless some new measurement is published (or an old

[2]The PDG does actually use some judgment in deciding whether to include a measurement or not.

measurement is dropped). Comparing the currently very well established value of the K^0_S mean lifetime with the early 1970's world average, we see that the 1970's value is off by more that 6σ (σ being the 1970's error-bar). The probability for obtaining a value this far off, or further, in either direction (due to a statistical fluctuation) is 2×10^{-9}.

Shockingly, the same general scenario is repeated 20 years later with the $B^\pm$ mean lifetime. (The 2002 PDG lifetime value, and the dotted line through it, are added here—for some reason the figure did not get updated in the most recent issue.) The 1992 $B^\pm$ mean lifetime is 8σ below the current value (very well established with the B-factories weighing in). The corresponding probability (or P-value) is 1.2×10^{-15}. (In both the K^0_S and $B^\pm$ (see [10]) cases there are excellent reasons for trusting the current value, so we do not compute the P-value under the alternative assumption that both the current world average and the suspect value are off by many σ (due to statistical fluctuations) from a true value that lies between the two. This is left as an exercise for the reader.)

Examining the $B^\pm$ mean lifetime plot in greater detail, one sees that the downward deviation from the current value grows from 0.9σ in 1986, to 2.5σ in 1988, to 4.5σ in 1990, to 8σ in 1992. But, within themselves, these values are quite consistent with each other—only after one has a good estimate of the true value is it evident that the world average $B^\pm$ mean lifetime is consistently biased low during this period. (Since the size of the error bars drops rapidly during this period, each point is dominated by new data.) It is quite likely that the experimenters during this period paid too much attention to the level of agreement between their new result and the measurements of the recent past. If one judges whether a result is ready for publication by its agreement with the current world average, such disasters can happen.

Blind analysis techniques can help reduce the likelihood of such problems. Blunders can always happen, but naturally when one sees an 8σ disagreement between one's current result and the world average, one goes back and checks[3] everything very carefully. On the other hand, when the new number is consistent with the old, there is a tendency to declare victory and move on. This asymmetry results in a statistical bias. A better procedure would be to list and schedule all the checks in advance of knowing the answer, and carry them out in either case (i.e. ask the question "What would we be checking if the answer were 8σ off?", write it down, and then do it in any case). Strictly

[3]See reference [11] for an excellent discussion of checks.

speaking, these checks are part of $d(X)$, since they are of the form "do this check, and if a mistake in $d(X)$ is discovered as a result, correct it".

5 Other Sources of Information

There are a number of recent references with information about blind analysis in high energy physics: A good place to start is the New York Times article by James Glanz[12]. Another source is the list of links in the blind analysis section of Aaron Roodman's home page[13]. Roodman is quoted in the NYT article, and he developed the blind analysis technique that was used in the BABAR $\sin 2\beta$ measurement. Especially interesting is Roodman's link to the "Draft Guidelines for Blind Analyses in BABAR".

Paul Harrison, also from BABAR, wrote an excellent review of blind analysis for the 2002 Durham Conference Proceedings that is available as reference [14]. It contains several additional historical examples, as well as a discussion of techniques.

The CDF B Group's "Good Analysis Practices" Committee[15] has a draft "Guidelines for Data Analysis" that explicitly discusses blind analysis, and offers much practical advice of a more general nature. Reading this document carefully, one discovers that some techniques that might fall under the purview of "blind analysis" are in fact recommended for all analyses—blind or otherwise. The collected "comments sent to the committee" are quite interesting as well.

Closely related is the section entitled "When is a Signal Significant?" in Byron Roe's *Probability and Statistics in Experimental Physics* [16], which discusses experimenter bias in bump hunting. Here, the experimenter searches for new resonances by trying many variations of the cuts (or looks in many channels)—the significance of any resulting "discovery" is likely to be overstated unless one can compensate somehow for the "false positives" that are likely to occur if enough histograms are examined.

6 Techniques

At most, we can only offer rather general advice; the experimenters who are doing the analysis must ultimately decide what is needed and what is doable. The techniques are, in many cases, trivial.

6.1 general principles

- It is not always possible to achieve perfect blindness in all cases. The classic example in the medical field is when the drug being tested has side effects with noticeable symptoms: the patients (and the doctors) may be able to tell whether the drug or the placebo is being administered. In such cases, one simply tries to achieve as much blindness as one can, rather than giving up completely.

- Blind analyses can suffer from most of the problems that arise in traditional analyses—blindness is not a panacea.

- In blind analysis, one does not have to plan everything before doing anything—it is only after the answer (or some related statistic) is known from real data that experimenter bias is possible. One is free to try out various plans (on simulation, or, in some cases, on real data with the answer hidden or shifted) and modify them or reject them before settling on the final plan.

- Blind analysis is intended to guard against experimenter bias, not fraud. The experimenters themselves carry out the blind procedures, which they have designed themselves to avoid knowing certain things until the appropriate time. Blind analysis is not intended to prevent them from cheating (i.e. pretending blindness while not practicing it). As in the past, the experimenters are assumed to be honest and trustworthy unless proven otherwise.

- Besides hiding the answer from the people actually performing the analysis, another benefit of blind analysis is that it also hides the answer from the physics group, godparents, advisers, and the collaboration in general. Clearly, anyone in a position to recommend alterations to the procedure, suggest checks, etc., is also a potential source of bias.

- Logically, one would hide the answer from the journal referees as well, but this would require a change in the typical journal's submission procedure—for example, having a "sanitized" version of the paper available for use by the referees, in addition to the normal version for publication. (Additionally, one might wish to hide the experiment's identity—CDF, DØ, BaBar, etc.—from the referees, but this seems hardly possible for high energy experiments.)

6.2 hiding the answer

The most basic technique is simply not printing out the answer. In some cases, certain subsidiary diagnostic plots also need to be re-centered to zero mean before examination. The classic example of this procedure is the hidden box method, which is the established procedure in the rare Kaon decay community. One hides the number of events in the signal region (i.e., the box) until the cuts have been finalized and the acceptance has been determined. At the final stage, the box is opened, and the answer (cross section measurement or limit) is computed.

6.3 shifting the answer

In some cases, it may be sufficient to shift the answer by adding a random (but fixed and unknown) offset Δ to the answer. For example, we suppose that the statistical uncertainty of a measurement of $\sin 2\beta$ is uncorrelated with the actual value of $\sin 2\beta$ (probably at least a good approximation). One then uses $\Delta + \sin 2\beta$ instead of $\sin 2\beta$ in the fitter's likelihood function, and the uncertainties are invariant, but the fitter's answer has a random offset of unknown size or direction.

Such a random, fixed and unknown offset can be produced by a single person using a hash function. A trivial example in C is:

```
#include <stdlib.h>
srand(3141591234); // set the seed for rand
const double delta = (10.0/RAND_MAX)*rand() - 5.0;
```

which gives a reproducible and fixed value for Δ that is a single random deviate uniform in the range $[-5, 5]$. (A defect of this example is that it is not system independent—it may give a different answer on Linux, SGI, etc.) Improved hashing schemes will undoubtedly occur to the reader after a little thought.

An advantage of this approach, vs simply not printing out the answer, is that it allows two independent groups to analyze the same real data and compare their answers—both having the same random offset—while the real answer remains unknown. Of course, access to the actual value of the seed used to produce the random offset should restricted to those executing the analysis—they are assumed to be trustworthy.

6.4 hiding (some of) the data

In this approach one performs a non-blind analysis on a subset of the data: the development dataset. For example, one might randomly split[4] all data event-by-event into two sets: A and B. The analysis procedure is developed using set A—set B is not looked at all. Once the analysis algorithm is finalized, set A is discarded, and the analysis is run on set B, which determines the final answer.

Unfortunately, in many cases, the experimenters don't have the luxury of having so much data that they can afford to throw away a significant fraction of it. This may lead them to publish the answer using both sets, which defeats the purpose of the procedure.

In addition, it is not clear that the procedure is always free from bias. For example, in the course of a non-blind analysis of set A, say a top mass measurement, the experimenters might need to do a detector calibration that will also be used for set B. If the calibration is unconsciously altered to bring the answer from set A closer to expectations, this bias will persist in the final answer determined from set B.

The method seems best suited to a case where many cut variations are tried on data in order to search for unanticipated signals (bump hunting being a prime example), but the analysis procedure is otherwise fixed. As discussed in [16], it is easy to be fooled by the statistical fluctuations that mimic new effects—if enough cut variations are investigated. In such cases, it is helpful to have the unexplored set B to confirm or refute any "discovery" in set A.

A much better procedure, when possible, is to "hide" all of the data, optimize the analysis entirely on Monte Carlo, and only look at the data after the analysis algorithm is frozen. But for some analyses, the Monte Carlo simulation does not model the data sufficiently well for this approach to be practical. And, of course, unanticipated new physics effects are not present in the Monte Carlo simulations.

[4] Often the data is split instead into "early" and "late" data sets, but this much less satisfactory.

7 Summary

Blind analysis techniques are recommended as a way to reduce the likelihood of experimenter bias, thereby also reducing the rate of wrong answers. The fundamental strategy is to avoid knowing the answer until the analysis procedure has been set. Since checks may lead to a change (or correction) of the procedure, they should be completed, or at least scheduled, before the answer is revealed. Despite the precautions, should a major (unanticipated, answer-changing) bug turn up after the answer is revealed, it must be explained in the publication, so that the readers can form their own opinion.

References

[1] J.A. Simpson and E.S.C. Weiner. "The Oxford English Dictionary", 2nd ed, (Oxford University Press, Oxford, 1989).

[2] A. Pais, "Inward Bound" (Oxford University Press, Oxford, 1986) p 363; E. Rutherford, Proc. Solvay Conference, (Gauthier-Villars, Paris 1934), p 177.

[3] George S. LaRue, James D. Phillips, and William M. Fairbank, Phys. Rev. Lett. **46**, (1981), p 967. `link.aps.org/abstract/PRL/v46/p967`

[4] P.F. Smith, Ann. Rev. Nucl. and Part Sci. **39**, (1989), p 73; L. Lyons, Phys. Reports **129**, (1985), p 225.

[5] J.D. Philips, W.M. Fairbank, and J. Navarro, Nucl. Instrum. Methods **A264**, (1988), p 125.

[6] H.B. Prosper, J.T. Linneman and W.A. Rolke, "A Glossary of Selected Statistical Terms" in Proceedings of the Conference on Advanced Techniques in Particle Physics, M. Whalley and L. Lyons (ed.), IPPP/02/39 (July 2002), p 314.
`www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/glossary.ps`

[7] W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet "Statistical Methods in Experimental Physics", (North-Holland Publishing Co, Amsterdam, 1971), §7.1.2, p 116.

[8] Harold Jeffreys, "Theory of Probability", 3rd ed., (Oxford University Press, Oxford, 1961), §5.63, p 308.

[9] K. Hagiwara et al., Phys. Rev. **D66**, 010001 (2002), p 15.
pdg.lbl.gov/2002/historyrpp.ps

[10] LEP B Lifetimes Working Group.
lepbosc.web.cern.ch/LEPBOSC/lifetimes/

[11] Roger Barlow, "Systematic Errors: Facts and Fictions", in Proceedings of the Conference on Advanced Techniques in Particle Physics, M. Whalley and L. Lyons (ed.), IPPP/02/39 (July 2002), p 134.
www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/barlow.ps

[12] James Glanz, "New Tactic in Physics: Hiding the Answer", New York Times, August 8, 2000.
www.nytimes.com/library/national/science/080800sci-particle-physics.html
www.jlab.org/news/internet/2000/new_tactic.html

[13] www.slac.stanford.edu/~roodman/babar.html#Blind

[14] Paul Harrison, "Blind Analyses" in Proceedings of the Conference on Advanced Techniques in Particle Physics, M. Whalley and L. Lyons (ed.), IPPP/02/39 (July 2002), p 278.
www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/harrison.ps

[15] www-cdf.fnal.gov/internal/physics/bottom/gap/gap.html

[16] Byron P. Roe, "Probability and Statistics in Experimental Physics", (Springer-Verlag, New York, 1992), §11.4, p 111.